A Toolkit for an Oral History Digital Archive [⋆]

Silvestre Lacerda[1]
lacerda@iantt.pt
Norberto Lopes[2], Nelma Moreira[2], Rogério Reis[2]
{nml,nam,rvr}@ncc.up.pt

[1] Direção-Geral de Arquivos/ Arquivo Nacional da Torre do Tombo
[2] DCC-FC& LIACC, Universidade do Porto

Abstract. In this work we propose an XML based toolkit for the construction of an oral history digital archive that allows the filing, classification and annotation of multimedia resources associated to a corpus of interviews. We describe the general organization of the archive and focus on content creation tools. In particular, we present a document editor for the classification and annotation of interview transcriptions that allows the definition of category hierarchies.

Keywords: XML languages, XML editors, oral history archives, multimodal resources, *metadata*, corpus

1 Introduction

In this work we propose an XML based toolkit for the construction of an oral history digital archive that allows the filing, classification and annotation of text and multimedia resources associated to a corpus of interviews.

The two basic objects used in this archive are interviews and persons. Each interview (to a person) can have several audio and video files, photos and other images attached, and several derived text documents that result from different analysis of the interview's data.

To ease the access, classification and search of the information, all text based documents should be in a semi-structured (XML) format and all multimedia documents should have *metadata* information associated. To help the construction of these documents, in particular by non computer specialists, dedicated software applications must be built. We present a document editor for the classification and annotation of interview transcriptions that allows the definition of category hierarchies.

This paper is organized as follows. In the next section we present the motivation and the scope of this work. Section 3 describes the general organization for the interview's archive. Section 4 presents a document editor for classification and annotation. Some related work is discussed in Section 5 and Section 6 concludes with some ongoing and future work.

[⋆] Work partially funded by FCT through the grant MTCIO (POCTI/CED/60786/2004) and POSI Program.

2 Motivation

This work is part of the research work developed by the *Documentation and Information Center on Working Class and Popular Movement of Porto* (**CDI**) of Universidade Popular do Porto [16] . **CDI**'s goals are the preservation of the social, cultural and political memory, and oral and social history of Porto, during the 20th century, as well as its diffusion. Available since 2001, **CDI**'s Web site contains information about several workers' organizations, including documental inventories, archival descriptions and digitalized documents; abstracts of workers' interviews and a chronology of workers' related events during the 20th century. Search in the site can be text based or using a special controlled vocabulary.

In what the oral history is concerning, **CDI** has already collected a corpus of about one hundred interviews with workers of different professions and with different social experiences. All biographical narratives were recorded in audio and video, and a transcription of the interview was produced. A small abstract of each narrative is already available in **CDI**'s Web site.

The quantity, diversity and quality of the collected information by the **CDI** inspires its study in a multidisciplinary approach. The research team of this project involves different social sciences domains (Linguistics, Education Sciences, History, Information Sciences), and Computer Science.

To ease the access, the search and the multi-disciplinary analysis of the biographical narratives, new software applications are being developed. In Silvestre *et. al* [9] we presented the main aims for an oral history archive and described some tools already implemented. We briefly review some of those tools in the next section. We would like to emphasize our commitment in using and developing *free software* tools and documents with open specifications, as the only way to ensure timeless accessibility, portability and easy updating. The choice of `XML` for the format of text documents illustrates well our options.

3 A Digital Interviews Archive

As we pointed out in the introduction, the two basic objects used in an oral history archive are interviews and persons. Each interview can have several associated resources. Here we will consider the following:

- audio and video recordings
- photos, digitalized documents, images
- text documents:
    - transcriptions
    - abstracts
    - classifications associated with structural content analysis
    - other documents

All documents must have *metadata* information associated. For the analysis and research the documents should have *annotations* that associate a concept to a text segment. A text segment can have multiple annotations. For a more efficient search, annotations should use controlled vocabularies, and specially *thesauri*. The *thesauri* will allow a more fine-grained classification of documents that can be of value for social scientists research.

In Silvestre *et. al* we presented a set of `XML` based software tools for the production, annotation and search of multimedia documents. In particular we specified `XML` languages for : *metadata* based on the *Dublin Core* standard [7]; *thesauri* based on the `Zthes` standard[19]; annotations based on a specific controlled vocabulary; transcriptions of interviews based on the Transcriber schema [12, 1, 2]; and, for small abstracts of the interviews contents organized in sections, called *stories*. We developed special purpose document editors for *metadata* and abstracts. Both for the *thesauri* and the abstracts an Web site was implemented that allows browsing and search. The `XML` schema used to define each language is `Relax NG` [17]. Besides its expressive power it has a very elegant (compact) syntax and a well-defined semantics based on regular tree languages that allows efficient and clear implementations.

The annotations language was easily extended to deal with other vocabularies. In the next section we will describe an extension of the abstract's editor to a new editor that allow general classification of transcriptions by the definition of arbitrary category hierarchies.

3.1 Organizing the Resources Associated with an Interview

To organize the resources associated with an interview we used an approach similar to the *IMDI* standard developed by ISLE (International Standard for Language Engineering) Meta-data Initiative [6, 3] for linguistic multimedia resources.

In the *IMDI* standard the main concept is *session* which *bundles all information about the circumstances and conditions of a linguistic event*. Although some of this information can be retrieved from the resources' *metadata*, other should be duplicated in each resource in order to be possible to group and collect it. For example, every document should have information about which interview it refers to and that would not be easy to infer without the notion of *interview*. We consider the main *metadata* elements proposed by *IMDI* and we specified the `session` `XML` element which the `Relax NG` schema is in Figure 1.

Note that we restricted the content model of many elements, as many linguistic specifications were not meaningful to our goals. Almost all elements are self-explanatory and similar to the ones defined in the *Dublin Core* standard. In the `content` element, the attribute `genre` describes the discourse type of the session contents. In our *corpus* the genre will be in general `interview`. The attribute `task` can be used to describe the topic of the session (for instance, *biographical narrative*).

The most relevant element for us is `person`. In Figure 2, the `Relax NG` schema for that element is given.

```
start = session
session = element session {
      attribute name {text},
      attribute title {text}?,
      attribute description {text}?,
      date+, place, project+, content,
      contributor*, resource*
}
place = element place {text}
project = element project {
   attribute name {text},
   attribute title {text}?,
   attribute description {text}?
   attribute person-id {text}?,
   contact?,
}
content = element content {
   attribute genre {text},
   attribute task {text}?,
   attribute description {text}?,
   subject*, language*
}
subject = element subject {text}
language = element language {text}
contributor = element contributor {
  attribute person-id  {text},
  attribute resource {text},
  attribute role {text}
}
resource = element resource {
         attribute type {text},
         attribute link {text},
         attribute access {"private" | "owner" | "public"}
}
```

Fig. 1. `Relax NG` schema for interviews.

Note again that, in contrast with the *IMDI* schema, the element `contributor` in a *session* has an attribute that refers to *person* and is not itself a `person` element. The value of the `role` attribute of the `contributor` element belongs to a controlled vocabulary that includes names as *interviewer*, *interviewee*, *transcriptor*, etc..

3.2 Building an Archive

An interviews' archive is a set of sessions, a set of persons and a set of associated resources. The organization of an archive can be specified as a tree of directories in a file system. The basic structure is presented in Figure 3.

```
start = person
person = element person {
   attribute id {text},
   attribute name {text},
   attribute fullname {text}?,
   attribute gender { "male" | "female" }?,
   birth? & situation? & language? & contact?,
   (photo | place | education | activity | organization | observation)*
   }
photo = element photo {text}
birth = element birth {date? & place?}
place = element place {text}
language = element language {text}
education = element education {text}
activity = element activity {
      attribute type {text}
      text }
situation = element situation {attribute type {text}},
   element session {attribute id {text}}*}
institution = element institution {text}
observation = element observation {text}
}
```

Fig. 2. `Relax NG` schema for persons.

The file `persons.xml` in the *directory* `_persondb` contains information about which persons are in the archive, each one described using a `person` element. In the same way, the file `sessions.xml` contains information about each session in the archive. These files will be manageable even if their size will reach some megabytes. This conclusion is based of some performance tests with more than 50000 records (sessions or persons). In this way we can profit from `XPath` facilities for querying and presenting the archive information.

There must exist a central archive. Each user can have a local archive which imports/exports information from and to the central archive. When a user exports a new version of an existing document a new revision is added, and the original document will not be modified.

4 Interviews Classification

For the data analysis of the biographical narratives corpus, the text annotations provide a useful indexation of the data, but is not enough for a structural content analysis. Some topics or categories can be defined globally driven by the research goals, but each narrative motivates the introduction of new categories. Those categories are then hierarchically organized into trees and several text segments can be associated to each category.

```
_archive/
      /_persondb/
               persons.xml
               /_photo
                 photo1.png
                 photo2.png
                  .
                  .
                  .
      /_sessiondb/
                sessions.xml
                E01/
                   trans1.xml
                   class1.xml
                   audio1.xml
                   video1.xml
                E02/
                  .
                  .
                  .
```

Fig. 3. Archive basic structure.

This content analysis can produce one or several classifications of the original transcription. Each classification can be used for several transcriptions and each transcription can be associated with several classifications.

A classification is then a forest (set of trees) where each node is labelled by a category and may have several text segments associated to. Currently, we only accept the association of a transcript to a classification. Figure 4 presents the `Relax NG` schema of a classification. The element `node` corresponds to a category. The element `content` corresponds to a text segment of a transcription and can be identified by a `start char` and an `end char`. This text segment (contained in the `value` element) cannot be modified, although it can have independent annotations in the classification and in the transcription. The attribute `link` can be used for identifying the transcription that segment of text belongs to (in the case that multiple transcripts are allowed).

In Figure 5 we present the new `XML` specification for annotations and for vocabularies (that are not a *thesaurus*).

In future implementations, each `content` should also be marked with insertions, deletions or substitutions. These would be useful for a printed version of a classification where some minor modifications of the original transcript can be allowed (for instance, for omitting punctuation or a private part of the discourse).

```
include "metadata.rnc"
start = classification
classification = element classification {
   metadata,
   name,
   description?,
   nodes?
}
nodes = element nodes { node+ }
node = element node {
   ref?,
   name,
   description?,
   comment?,
   contents?,
   nodes?
}
contents = element contents { content+ }
content = element content {
   attribute startchar { text },
   attribute endchar { text },
   attribute  link  { text }?,
   ref?,
   description?,
   comment?,
   value
}
name = element name { text }
description = element description { text }
comment = element comment { text }
ref = attribute ref { text }
value = element value { text }
```

Fig. 4. The `Relax NG` schema for classifications.

4.1 The Classification Editor

The interviews' transcriptions are `XML` documents with a specification based on the Transcriber schema [12, 1, 2], but supporting *metadata* and annotations. Its `Relax NG` schema is presented in Figure 7. For now we do not have a transcriber application, but only some converters from several formats into our transcription format.

For the classification of transcriptions we developed an editor that allows the dynamic definition of categories trees and builts new documents based on the information in the transcriptions.

Figure 6 presents a screen-shot of the editor. The editor main frame is divided into two panels. In the left panel, an hierarchy of categories can be built. A new node can be created and attached anywhere. A node has a `name` (the category), a

```
annotations = element annotations {
   attribute vocabulary {text},
   attribute type {text},
   attribute marks? { text },
   annotation+
}
annotation = element annotation {
   attribute startchar { text },
   attribute endchar { text },
   attribute type { text },
   attribute normtext { text },
   text
}
```

```
vocabulary = element vocabulary {
      attribute name { text },
      attribute date { text },
      attribute link { text}
      description?,
      entry+
}
description= element description {
   attribute lang {text}
   text
}
entry= element entry {
   attribute value {text}
   text
}
```

Fig. 5. The Relax NG schemas for annotations and vocabularies, respectively.

description and a comment. There may exists several top nodes, so several trees are allowed. It is also possible to change a node's place or to edit its attributes.

The right panel, has two windows: the lower one for a transcription and the upper one, will have the contents associated with a category. In the transcription, the different speakers are identified by a number. The identity of each speaker can be checked by selecting an option in the View menu.

To associate a text segment to a category, choose the category and just select the text segment from the transcription. Then, the text segment will appear in the upper window, and a small description will be also attached below it's category (in the left side). In Figure 6 the highlighted text is a description of the content which is shown in the upper window on the right, and which category is *Primaria*.

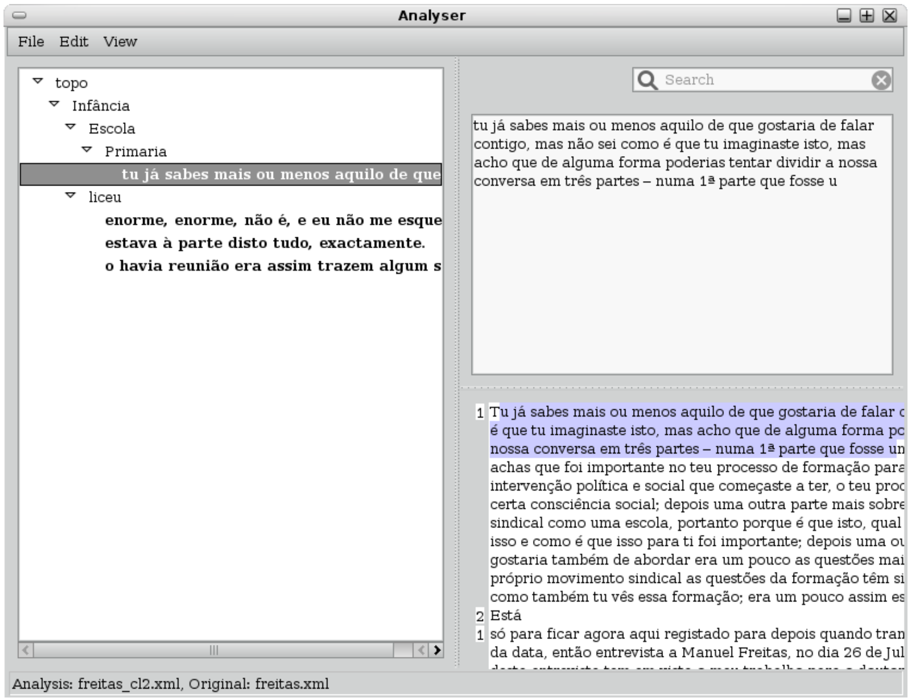
When a category is selected all its contents are listed in the upper window on the right and each text segment is highlighted in the transcription (lower window on the right).

The options in the Edit menu allow the edition of *metadata* and of some preferences.

The options in the File menu allow to open/create an archive and choose a transcription (we plan to allow choosing more than one transcription simultaneously). It is possible then to create a new classification or open an existing one. A classification can be saved or exported to an HTML file. The category hierarchy can be saved independently (what we call a schema) in order to reuse it for other classification (and possibly for other transcriptions).

4.2 Implementation

The main programming language of this project is Python [18].



topo
Infância
Escola
Primaria
tu já sabes mais ou menos aquilo de que
liceu
enorme, enorme, não é, e eu não me esque
estava à parte disto tudo, exactamente.
o havia reunião era assim trazem algum s

tu já sabes mais ou menos aquilo de que gostaria de falar
contigo, mas não sei como é que tu imaginaste isto, mas
acho que de alguma forma poderias tentar dividir a nossa
conversa em três partes – numa 1ª parte que fosse u

1 Tu já sabes mais ou menos aquilo de que gostaria de falar c
é que tu imaginaste isto, mas acho que de alguma forma po
nossa conversa em três partes – numa 1ª parte que fosse un
achas que foi importante no teu processo de formação para
intervenção política e social que começaste a ter, o teu proc
certa consciência social; depois uma outra parte mais sobre
sindical como uma escola, portanto porque é que isto, qual
isso e como é que isso para ti foi importante; depois uma ou
gostaria também de abordar era um pouco as questões mai
próprio movimento sindical as questões da formação têm si
como também tu vês essa formação; era um pouco assim es
2 Está
1 só para ficar agora aqui registado para depois quando tran
da data, então entrevista a Manuel Freitas, no dia 26 de Jul

Analysis: freitas_cl2.xml, Original: freitas.xml

Fig. 6. The classification editor.

The graphical interfaces are implemented with the `wxWidgets` `API` [15]. Text editor objects are based in the `wxStyledTextCrl` class, that implements the editing component `Scintilla` [5]. The category tree graphical interface uses the `CustomTreeCtrl` class.

`XML` processing uses the `lxml` [11] library that implements the `elementtree` `API` [10], but with greater support for `XSLT` and `XPath`. This `API` is `Python` based and much faster and less space consuming than `DOM` or `SAX` `Python` implementations.

For the classification editor, the classification's tree structure is built using `Python` classes and it is saved as a document in the classification `XML` language.

5 Related Qualitative Data Analysis tools

There are several software tools available for the qualitative data analysis of a set of documents, that can be used for the structural content analysis of our corpus of interviews. From this set, we can refer the *Ethnograph* [14], the `Atlas.ti` project [4, 13], and the *Nvivo 7* [8]. All those tools are commercial proprietary software products with very expensive licenses. Only the *Atlas.ti* project can use `XML` files.

```
start = transcript
transcript = element transcript {metadata, actor+, section*, annotations*}
section = element section {
      attribute title { text },
      attribute desc { text },
      attribute starttime { text }?,
      attribute endtime { text }?,
      conversation+
   }
conversation = element conversation {
      attribute title { text },
      attribute actor-id { text },
      attribute starttime { text }?,
      attribute endtime { text }?,
      (comment | event | sync | text)+
   }
comment = element Comment {
   attribute desc {text}
}
sync = element sync {
   attribute time {text}
}
event = element event {
   attribute desc {text}
   attribute type {text}
   attribute extend {text}
}
```

Fig. 7. `Relax NG` schema for transcriptions.

The *Nvivo 7* is the most popular in the academic world. Its previous versions were not very user friendly and it was difficult to be used by a non computer specialist. *Nvivo 7* allows to create a project with several documents (in the RTF format). Arbitrary (free) nodes can be created and text segments associated to those nodes. The nodes can then be organized in a tree. Search facilities includes search for nodes or text, and proximity search of two items.

In comparison, our approach includes the basic features of this system (for now, for a single document) with the advantages of being free software and of using semi-structured documents with open formats.

6 Conclusion

In this paper we described ongoing work towards the construction of an oral history digital archive based on a corpus of interviews. We also described an editor that aims to help social scientists in the corpus structural content analysis. The current version of the editor works only with one interview (transcription) at a time. This is not a major drawnback as in most of the cases each interview has its own set of category hierarchies. Although it is still in development, it is already being used by some project team members. We are now extending the editor to allow the classification of multiple interviews simultaneously and improving its search facilities. For that it is essential to have a precise notion of an *archive*. In this paper, we also described the basic structure for building an oral history archive. This organization is also essential for an improved search in the corpus. We plan to implement search mechanisms that will allow

- querying the archive using controlled vocabularies and in special *thesauri*, or more complex ontologies;
- more complex queries, by the development of a specific query language fitted to the archive contents and simple to use;
- save the query results in a file for future use.

In what dissemination is concerned we note that provision must be made in order to restrict the access to several documents of the archive although general sessions information and documents *metadata* should be in general accessible.

7 Acknowledgements

We thank several members of Universidade Popular do Porto and team members of the project *Memories of Work: Construction Processes of a Worker's Identity* for their comments and suggestions. Specially we would like to thank Cristina Nogueira and Teresa Medina for many discussions about the classification editor. Thanks are also due to the referees whose comments and suggestions allowed to improve a first version of this paper.

References

1. Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. Transcriber: a free tool for segmenting, labeling and transcribing speech. In *First International Conference on Language Resources and Evaluation (LREC)*, pages 1373–1376, 1998.
2. Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1–2):5–2, 2001.
3. D. G. Broeder, H. Brugman, A. Russel, and P. Wittenburg. A browsable corpus: accessing linguistic resources the easy way. In *LREC 2000 Workshop*, Athens, 2000.
4. ATLAS.ti Scientific Software Development. ATLAS.ti. `http://www.atlasti.com/`, Date of Access: October 2007.
5. Scintilla Project Group. Scintilla. `http://www.scintilla.org/`, Date of Access:2007.
6. IMDI Team. IMDI Metadata Elements for Session Descriptions. Technical report, MPI Nijmegen, October 2003.

7. The Dublin Core Metadata Initiative. Dublin Core Metadata Terms. http://dublincore.org/, Date of Access: October 2007.
8. QSR International. Nvivo. http://www.qsrinternational.com/, Date of Access: November 2007.
9. Silvestre Lacerda, Norberto Lopes, Nelma Moreira, and Rogério Reis. Ferramentas para a construção de arquivos digitais de história oral. In Luís Carriço José Carlos Ramalho, João Correia Lopes, editor, *Actas XATA 2007, XML: aplicações e tecnologias associadas*, pages 139–150. Universidade de Lisboa, Fevereiro 2007.
10. Fredrik Lundh. ElementTree API. http://effbot.org/zone/element-index.html, Date of Access:2007.
11. lxml development team. lxml. http://codespeak.net/lxml/, Date of Access:2007.
12. Sylvain Galliano Mathieu Manta, Fabien Antoine and Claude Barras. Transcriber. http://trans.sourceforge.net/, Date of Access: October 2007.
13. Thomas Muhr. Increasing the reusability of qualitative data with xml. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, Date of Access: October 2007.
14. Qualis Research. The Ethnograph. http://www.QualisResearch.com, Date of Access: October 2007.
15. Julian Smart, Robert Roebling, Vadim Zeitlin, and Robin Dunn. *wxWidgets 2.6.3: A portable C++ and Python GUI toolkit.*
16. Universidade Popular do Porto. Centro de Documentação e Informação sobre o Movimento Operário e Popular do Porto. http://cdi.upp.pt/.
17. Eric van der Vlist. *RELAX NG.* O'Reilly, 2003.
18. Guido van Rossum. *Python Library Reference*, 2.4.2 edition, 2005.
19. The Zthes working group. The Zthes specifications for thesaurus representation, access and navigation. http://zthes.z3950.org/, Date of Access: October 2007.